MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231139

Speech Emotion Recognition Applied to Real-World Medical Consultation

Ching-Tzu HUANG[a,b], Chih-Wei HUANG[b], Hsuan-Chia YANG[a,b,c] and Yu-Chuan (Jack) LI[a,b,d,1]
[a] *Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan*
[b] *International Center for Health Information and Technology(ICHIT), Taipei Medical University, Taipei, Taiwan*
[c] *Graduate Institute of Data Science, College of Management, Taipei Medical University, Taipei, Taiwan*
[d] *Department of Dermatology, Wanfang Hospital, Taipei Medical University, Taipei, Taiwan*
ORCiD ID: HC Yang https://orcid.org/0000-0001-9198-0697, YC Jack Li https://orcid.org/0000-0001-6497-4232, CW Huang https://orcid.org/0000-0002-2551-6199

Abstract. Since 2020, the COVID-19 epidemic has changed our lives in healthcare behaviors. Forced to wear masks influenced doctor-patient interaction perceptions truly, thus, to build a satisfying relationship is not just empathize with facial expressions. The voice becomes more important for the sake of conquering the burden of masks. Hence, verbal and non-verbal communication will be crucial criteria for doctor-patient interaction during medical consultations and other conversations. In these years, speech emotion recognition has been a popular research domain. In spite of abundant work conducted, nonverbal emotion recognition in medical scenarios is still required to reveal. In this study, we investigate YAMNet transfer learning on Chinese Mandarin speech corpus NTHU-NTUA Chinese Interactive Emotion Corpus (NNIME) and use real-world dermatology clinic recording to test the generalization capability. The results showed that the accuracy validated on NNIME data was 0.59 for activation prediction and 0.57 for valence. Furthermore, the validation accuracy on the doctor-patient dataset was 0.24 for activation and 0.58 for valence, respectively.

Keywords. Speech emotion recognition, medical education, doctor-patient communication, YAMNet transfer learning, bidirectional long short-term memory networks

1. Introduction

According to the research, doctor-patient relationships play an important role in healthcare services satisfaction and compliance [1]. A satisfying cognitive interaction led to increased physician trust and expertise perception [2]. However, masks cover a part of our most significant way to express empathy during some particular situations,

[1] Corresponding Author: Yu-Chuan (Jack) Li, M.D., Ph.D. email: jack@tmu.edu.tw

e.g. epidemics [3]. The empathic impression through vocals becomes necessary. Many studies already showed that artificial neural network could predict emotion categories over different non-verbal measurements. Automatic speech emotion recognition (SER) is one of the essential technologies for estimating healthcare service consultation quality, which has been developed for years for various life applications. Notwithstanding, the real-world practical application of SER in the medical field is still insufficient [4].

In this case, we explored the generalization of speech emotion recognition models on real clinical scenario recordings. The goal is to develop a deep learning model specialized in doctor-patient communication.

2. Methods

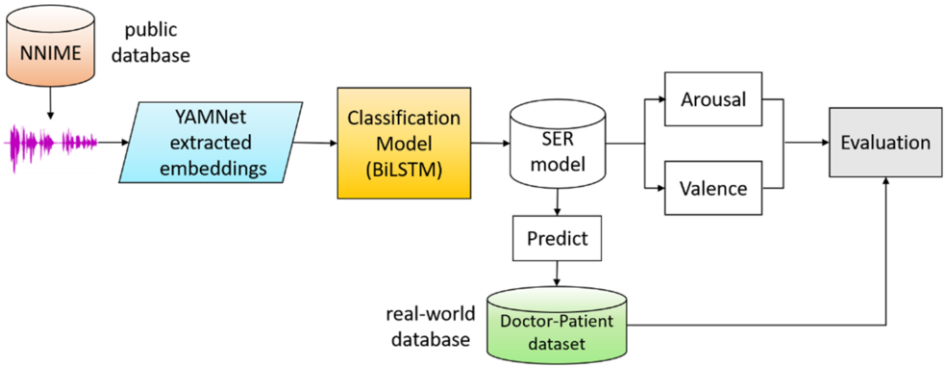

Figure 1. The development workflow of the model.

2.1. Data sets

2.1.1. NTHU-NTUA Chinese Interactive Emotion Corpus

We used the NTHU-NTUA Chinese Interactive Emotion Corpus (NNIME) to run deep neural network training on the two primitive emotion patterns, activation, and valence. The original sentence-level annotation by naïve raters coverage is 1 to 5, to do initial experiment we modified it to 1 to 3. (1 represents negative emotion, 3 represents positive for valence; 1 represents calm, 3 represents excited for activation)

The dataset contains a total of 44 subjects who had prior real-life experience in professional acting(22 females, 20 males). The age ranged from 19 to 30. All of the recordings are in Mandarin Chinese. The collections are natural affective dyadic interactions. To increase the ability of generalization, we used all the sentence-level data(4806 audio segments) for training and testing [5].

2.1.2. Doctor-Patient interaction Corpus

Our team collected the real-practice video recording from the dermatology clinic room in the Wanfang Hospital, Taipei, Taiwan. We obtained 378 different patients with 4 doctors(2 females, 2 males). In this study, we selected 11 recordings from the database by sex ratio of the doctors. There are 869 sentence-level audio segments after manual segmentation. The annotation labels are the same as the NNIME dataset, valence, and activation levels 1 to 3 [6].

2.1.3. YAMNet

YAMNet is a deep neural network feature extractor that was trained by AudioSet corpus, which contains over 2,000,000 video segments of 512 type audio events and uses the MobileNet v1 networks architecture [7].

2.1.4. Bi-directional Long Short-Term Memory Networks

To predict the time series emotion features, Bi-directional Long Short-Term Memory Networks(BiLSTM) were used in this study. A Bidirectional LSTM, the architecture consists of reversed Long Short-Term Memory Networks [8]. In the previous study, BiLSTM can improve speech emotion recognition task accuracy, providing more completed contextual information of acoustic features [9].

2.1.5. Development of the model

In the preprocessing procedure, the emotion sentences were segmented by an approach that set two seconds a chunk, and the overlapping part is one second, as shown in Figure 2. We resampled the audio files to 16 kHz mono, then extracted the 1024-dimension embeddings from YAMNet architecture. After extracting from the pre-trained YAMNet model, the features were fed to a deep neural network we created. The network contains two BiLSTM layers. The model was optimized using Adam optimizer [10], while the Softmax activation function was used in the output layer. The dropout value of 0.2 was used to prevent overfitting and was applied in the hidden layers [11].

Figure 2 shows the neural network structure of the study. The NNIME dataset was split into the training set, test set, and validation set by the proportion 72%, 10%, and 18%.

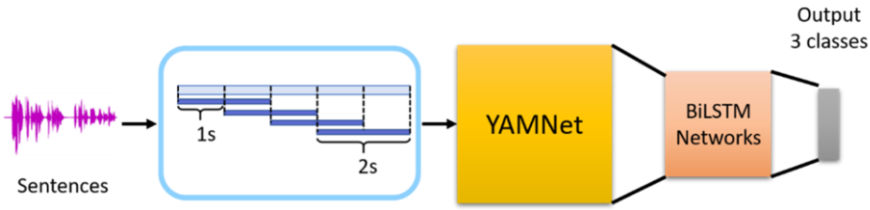

Figure 2:Deep Neural Network Training Architecture

3. Results

Table 1 shows the result of the validation on the NNIME. The performance of the model was observed with an unweighted accuracy of 0.58 and the weighted average F1-score of 0.68 in the valence task; for activation, the unweighted accuracy achieves 0.59, the F1-score performs only 0.6. Moreover, Figure 3 depicts the model prediction on the real doctor-patient conversation dataset. The valence model performance concentrates on neutral emotion. Otherwise, the activation model accuracy is much lower than the valence model. When the actual label is high activation, the model predicted lower.

Table 1. The result of the model validation on the NNIME dataset

		Precision	Recall	F1-score	Unweighted accuracy
Activation	**Macro average**	0.59	0.61	0.58	0.59
	Weighted average	0.67	0.59	0.60	
Valence	**Macro average**	0.33	0.30	0.28	0.58
	Weighted average	0.83	0.58	0.68	

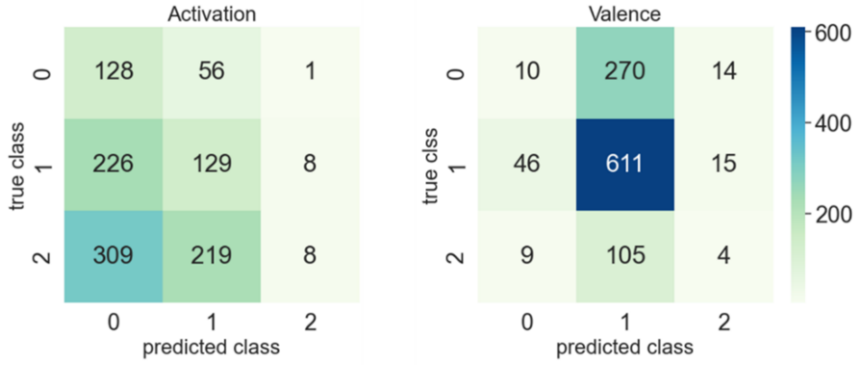

Figure 3. The confusion matrices with the number of predictions in the doctor-patient interaction dataset.

4. Discussion

Our study findings indicate that the acting speech emotion dataset has the potential on applying in real doctor-patient conversation corpus. However, the accuracy of the activation recognition task is relatively low. This disclosure suggests we can attempt to experiment with various methodologies to decrease the disparity between acting speech data and real-world speech data, such as domain-invariant feature extraction based on the public speech emotion dataset as an alternative or building a new multi-task model specialized on our doctor-patient dataset. Although using pre-trained network YAMNet can reduce data complexity to increase the efficiency of emotion pattern recognition, it may also bring disadvantages to the pattern recognition since simplifying some key features of the speech data.

5. Conclusions

In this study, we hypothesized the acting emotion speech data can adapt the real-world doctor-patient conversation. Therefore a speech emotion recognition model using transfer learning was developed to predict the two primitive emotion patterns(valence, and activation). Our model showed the performance is better in the same dataset than the external dataset, possibly by the reason of different feature distribution. Nevertheless, our doctor patient dataset still holds advantages to develop an automatic model to help trained medical students to be aware of their emotional reactions when practicing non-verbal communication except for learning verbal professional knowledge. In the future, more methodologies are required to be investigated to resolve the cross-corpus generalization difficulty. Further experiments should focus on the various feature

extraction and different methodologies, such as other temporal models as well as convolutional neural networks using the spectral or prosody features. ,For the reason that the real world data is not easily available, cross-corpus methods also should be considered.

Acknowledgements

This research was funded by the National Science and Technology Council (grant number: NSTC 110-2221-E-038 -002-MY2 and NSTC 110-2320-B-038-029-MY3) and the Ministry of Education in Taiwan.

References

[1] Kim SS, Kaplowitz S, Johnston MV. The effects of physician empathy on patient satisfaction and compliance. Eval Health Prof. 2004 Sep;27(3):237-51, doi: 10.1177/0163278704267037.
[2] Riess H, Kraft-Todd G. E.M.P.A.T.H.Y.: a tool to enhance nonverbal communication between clinicians and their patients. Acad Med. 2014 Aug;89(8):1108-12, doi: 10.1097/ACM.0000000000000287.
[3] Grundmann F, Epstude K, Scheibe S. Face masks reduce emotion-recognition accuracy and perceived closeness. PLoS One. 2021 Apr;16(4):e0249792, doi: 10.1371/journal.pone.0249792.
[4] Abbaschian BJ, Sierra-Sosa D, Elmaghraby A. Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. Sensors (Basel). 2021 Feb;21(4):1249, doi: 10.3390/s21041249.
[5] Chou HC, Lin WC, Chang LC, Li CC, Ma HP, Lee CC. NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus. In2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). 2017 Oct 23 (pp. 292-298). IEEE, doi: 10.1109/ACII.2017.8273615.
[6] Cowie R, Cornelius RR. Describing the emotional states that are expressed in speech. Speech communication. 2003 Apr;40(1-2):5-32, doi: 10.1016/S0167-6393(02)00071-7.
[7] M. Plakal and D. Ellis, "Yamnet," https://github.com/tensorflow/models/tree/master/research/audioset/yamnet, 2020.
[8] Graves A. Long Short-Term Memory. In: Supervised Sequence Labelling with Recurrent Neural Networks. Studies in Computational Intelligence, vol 385. Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-24797-2_4
[9] Meng H, Yan T, Yuan F, Wei H. Speech emotion recognition from 3D log-mel spectrograms with deep learning network. IEEE access. 2019 Aug;7:125868-81, doi: 10.1109/ACCESS.2019.2938007.
[10] Kingma DP. ba J. Adam: A method for stochastic optimization. 2014 Dec;1412:80, doi: 10.48550/arXiv.1412.6980.
[11] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research. 2014 Jan;15(1):1929-58, doi: 10.5555/2627435.2670313.
[12] Zhang S, Liu R, Tao X, Zhao X. Deep Cross-Corpus Speech Emotion Recognition: Recent Advances and Perspectives. Front Neurorobot. 2021 Nov;15:784514, doi: 10.3389/fnbot.2021.784514.